Proc. Natl. Acad. Sci. USA
Vol. 84, pp. 5625–5629, August 1987
Biochemistry

The G glycoprotein of human respiratory syncytial viruses of subgroups A and B: Extensive sequence divergence between antigenically related proteins

(viral attachment protein/O-linked glycosylation/protein structure)

PHILIP R. JOHNSON*†, MELANIE K. SPRIGGS*, ROBERT A. OLMSTED*, AND PETER L. COLLINS*

*Laboratory of Infectious Diseases, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, MD 20892; and †Department of Pediatrics, Vanderbilt University, Nashville, TN 37232

Communicated by Robert M. Chanock, April 28, 1987

ABSTRACT **Two major antigenic subgroups (designated A and B) have been described for human respiratory syncytial virus (RSV). Previously, on the basis of reactivity patterns with monoclonal antibodies, the greatest intersubgroup variation was shown to occur in the G protein, the putative attachment glycoprotein. To delineate the molecular basis for this variation, we have determined the nucleotide and deduced amino acid sequences of the G mRNAs and proteins representing a subgroup A (Long strain) and a subgroup B (18537 strain) virus. These sequences were compared to the available G mRNA sequence for another subgroup A (A2 strain) virus. The Long G protein shared 94% amino acid identity with the A2 G protein. In contrast, the 18537 G protein shared only 53% amino acid identity with the A2 sequence; interestingly, most of the sequence divergence occurred in the proposed extracellular domain of the G protein. This extensive divergence for the G protein was significantly greater than that observed for other RSV proteins. Despite this considerable divarication, the proposed extracellular domains of the G proteins contained a single region of highly conserved sequence and secondary structure that may represent a conserved structural or function domain, perhaps involved in attachment to cellular receptors. Furthermore, this conserved region may comprise part of an epitope that is shared between the two subgroup G proteins and may significantly contribute to the fact that, despite extensive overall amino acid sequence divergence, the RSV G proteins maintain significant antigenic relatedness.**

Human respiratory syncytial virus (RSV) is the most important infectious agent of acute lower respiratory tract illness in infants and young children (1). RSV shares general characteristics with and belongs to the Paramyxoviridae family (genus *Pneumovirus*) but differs significantly from other members of the family in its genetic map, gene products, and intergenic sequences (2). Two antigenic subgroups (designated A and B) have been described for RSV based on *in vitro* neutralization assays and reactivity patterns with monoclonal antibodies (3–5). Panels of monoclonal antibodies have discriminated subgroup-specific changes in at least six viral proteins (N, P, M, G, F, and 22K); interestingly, the greatest variation was in G, the putative attachment protein (4, 5). The G protein induces neutralizing antibodies against RSV in animals (6–8) and appears to account for the subgroup specificity of *in vitro* neutralization assays (P.R.J., R.A.O., P.L.C., G. A. Prince, B. R. Murphy, D. W. Alling, and E. E. Walsh, unpublished data). With regard to vaccine strategies for RSV, it was important to determine whether the specificities of monoclonal and neutralizing antibodies were due to extensive sequence divergence or to variation in a limited set of antigenic sites.

Complete sequences were available for 9 of the 10 mRNAs, including the G mRNA, of a subgroup A strain A2 (ref. 2 and references cited therein). To define the extent of sequence divergence between and within the two RSV subgroups, we undertook cDNA cloning and sequencing of a subgroup B strain (18537), as well as a second subgroup A strain (Long). In this report, we describe comparative analysis of the nucleotide and deduced amino acid sequences for the G mRNAs and proteins of the 18537, Long, and A2 strains. The G protein was found to have extensive sequence variability between subgroups, compared with the substantially less extensive divergence observed for other proteins. However, despite the overall G protein diversity between subgroups, regions existed that were highly conserved and were candidates to be domains important in transport, processing, biological activity, and cross-subgroup neutralization.

MATERIALS AND METHODS

cDNA Synthesis and Cloning. Double-stranded cDNA was synthesized by the method of Gubler and Hoffman (9) using RSV-infected cell mRNA as template, inserted into the *Sma* I site of the plasmid vector pTZ18R (Pharmacia), and cloned into *Escherichia coli* strain HB101.

Nucleotide Sequencing. Dideoxynucleotide sequencing of recombinant plasmids and RSV mRNA, using synthetic DNA primers, was performed as described (10, 11).

RESULTS

cDNA Cloning and Nucleotide Sequencing. cDNA libraries were constructed using as template mRNA from cells infected with RSV strains Long (subgroup A) or 18537 (subgroup B). Initially, Long and 18537 libraries were screened by heterologous hybridization with RSV strain A2 (subgroup A) cDNA probes representing the N, F, and G mRNAs. This identified corresponding Long and 18537 cDNAs, except that 18537 G cDNAs were not identified unambiguously. One possible explanation was that the sequence of the 18537 G cDNA was not sufficiently related to allow unambiguous identification by heterologous hybridization, and this was observed to be the case as indicated by the results described below.

As an alternative approach, viral-specific cDNAs in the 18537 library were detected by differential hybridization with radiolabeled cDNAs of mRNAs from uninfected and RSV 18537-infected cells. The 18537-specific cDNAs were then analyzed by nucleotide sequencing. Clones representing nearly complete cDNA copies of 9 of the 10 18537 mRNAs

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviation: RSV, respiratory syncytial virus.

(excluding L) were identified by nucleotide sequence homology (67–86%; see below and unpublished data) with the published A2 mRNA sequences.

The sequence of the 18537 G mRNA [921 nucleotides exclusive of poly(A); Fig. 1] was determined from four overlapping cDNA clones: pDD66 (representing nucleotides 18–921 of the complete mRNA sequence), pK38 (nucleotides 224–921), pV63 (nucleotides 394–921), and pG9 (nucleotides 18–921). All four inserts were coterminal at one end, each followed by poly(A) tracts of various lengths (6–30 nucleotides), thereby identifying the 3′ end of the 18537 G mRNA. The exact 5′ end of the mRNA was mapped by primer extension with a synthetic oligonucleotide (complementary to the mRNA from nucleotides 79–99) under conditions for dideoxynucleotide sequencing (data not shown); however, the identity of the first 6 nucleotides could not be determined unambiguously because of strong stop bands in each lane.

The sequence of the Long mRNA (Fig. 1) was determined from three overlapping cDNAs: pLF63 (nucleotides 9–464), pLB69 (nucleotides 409–917), and pA3 (nucleotides 596–917). The latter two cDNAs ended with poly(A) tracts of 10 and 11 bases, respectively, thus identifying the 3′ end of the Long G mRNA. The exact 5′ end of the Long G mRNA was not determined.

Features of the Nucleotide Sequences. The nucleotide sequence of the Long G mRNA was highly homologous with the A2 sequence, sharing 96% nucleotide identity (Fig. 1). A single nucleotide deletion, confirmed in two independent cDNAs, was observed at the 3′ end next to the poly(A) tract. In contrast, the 18537 G mRNA shared only 67% nucleotide identity with the A2 G mRNA. The overall lengths (921 vs. 918 nucleotides) were similar; an in-frame 3-nucleotide insertion occurred after position 642. As was previously noted for A2 (13, 14), the second ATG codon (positions 16–18) began the major open reading frame. However, the termination codon for the 18537 mRNA began 21 nucleotides (position 889 vs. 910) earlier than the A2 mRNA termination codon and was confirmed in four independent cDNAs. Thus, the 18537 mRNA sequence was predicted to encode a protein of 292 amino acids, 6 shorter than the 298 amino acids predicted for the A2 and Long mRNAs.

Features of the RSV G Glycoprotein. The predicted amino acid sequence of the Long G protein shared extensive homology with the A2 sequence (94% identity; Fig. 2). In contrast, the 18537 G protein sequence shared only 53% sequence identity. Somewhat surprisingly, this was substantially less than the nucleotide homology (67%) and was due to a high degree of nucleotide substitutions in first and second positions of codon triplets. Many of the amino acid substitutions appeared to be nonconservative, based on the scoring matrix of Dayhoff (15) (data not shown).

Previous work suggested that >50% of the molecular weight of the mature A2 G glycoprotein, as estimated by $NaDodSO_4$/PAGE, was contributed by carbohydrate, mainly in the form of O-linked oligosaccharides (13, 14). Potential acceptor sites for O-linked sugars, serine and threonine, comprised 31% of the total amino acid composition of the A2 G protein. Although the positions of these residues appeared to be randomly substituted (only 40% of the total serine and threonine residues occurred at the same positions in the two sequences), in the 18537 G protein a high content of serine and threonine (29%) was maintained. Therefore, the overall content, rather than the exact position of potential acceptor sites, was a conserved feature of the 18537 G protein. Since O-linked glycosylation has been noted to have a disproportionate effect on the relative mobility of glycoproteins in $NaDodSO_4$/PAGE (16), the actual contribution of O-linked carbohydrate to the molecular weight of the mature RSV G protein might be less than previously assumed and remains to be assayed directly.

A relatively smaller contribution of N-linked oligosaccharides to the electrophoretic mobility of the A2 G protein has

```
LONG          X                                                C      A                                                G
  A2  GGGGCAAATGCAAACATGTCCAAAAACAAGGACCAACGCACCGCTAAGACATTAGAAAGGACCTGGGACACTCTCAATCATTTATTATTCATATCATCGTGCTTA 105
18537 NNNNNN       C          C     A T       T C G   TC T   A        T     T      C A TG A     C T T

LONG                        A                                                          A
  A2  TATAAGTTAAATCTTAAATCTGTAGCACAAATCACATTATCCATTCTGGCAATGATAATCTCAACTTCACTTATAATTGCAGCCATCATATTCATAGCCTCGGCA 210
18537   C GA      T A      A          AG  C G  AG  T                    C  T  C             A        CAT   T  C

LONG                  T                                                                  T  G                  C
  A2  AACCACAAAGTCACACCAACAACTGCAATCATACAAGATGCAACAAGCCAGATCAAGAACACAACCCCAACATACCTCACCCAGAATCCTCAGCTTGGAATCAGT 315
18537   T        T    T      G TT CAG T   ACAAT  A  A   C CTG A  A AC T T C  C     T  T  AGTC  A CAGAAA GG   AC

LONG  TT   C    T                      C                                          A       C               T
  A2  CCCTCTAATCCGTCTGAAATTACATCACAAATCACCACCATACTAGCTTCAACAACACCAGGAGTCAAGTCAACCCTGCAATCCACAACAGTCAAGACCAAAAAC 420
18537 T A  C  A AAC CAC  CC       C    CA   A ATTC   CA   T T     AATACA  A   GAAACA  CCAT       CAC A      GG

LONG           C                        T                                  A                          C       T
  A2  ACAACAACAACTCAAACACAACCCAGCAAGCCCACCACAAAACAACGCCAAAACAAACCACCAAGCAAACCCAATAATGATTTTCACTTTGAAGTGTTCAACTTT 525
18537  G  TC  C   TC      GA   A      A G       TC   TTC   A  T       AA     A  AG      AC  T              T

LONG                                                                                          C       T
  A2  GTACCCTGCAGCATATGCAGCAACAATCCAACCTGCTGGGCTATCTGCAAAAGAATACCAAACAAAAAACCAGGAAAGAAAACCACTACCAAGCCCACAAAAAAA 630
18537   T     T  T     TG T  T    A CT    AAAT C          C       G   C      AAG  A   C A  C T   A        C

LONG       T     ---                 A                  C
  A2  CCAACCCTCAAG---ACAACCAAAAAAGATCCCAAACCTCAAACCACTAAATCAAAGGAAGTACCCACCACCAAGCCCACAGAAGAGCCAACCATCAACACCACC 732
18537       AC   AACC    A     G   C      A A C G   AA TGC    AA   A AT  T      C  AG  A  A A      C   G     A 735

LONG                              AA                  A                                  C                T
  A2  AAAACAAACATCATAACTACACTACTCACCTCCAACACCACAGGAAATCCAGAACTCACAAGTCAAATGGAAACCTTCCACTCAACTTCCTCCGAAGGCAATCCA 837
18537  G   G G   C  GC T T   A TC    GTGCT GA    ATC C    A  TA     TC   CA C  T  C          CA        AA  CA  C 840

LONG                  C             C                 A                T -  917
  A2  AGCCCTTCTCAAGTCTCTACAACATCCGAGTACCCATCACAACCTTCATCTCCACCCAACACACCACGCCAGTAGTTACTT  918
18537    T CA A   A AC C   G        CC T CA  TT AA C TAA TA AAA C TAGT   ATGCT      T C  921
```

FIG. 1. Alignment of nucleotide sequences (5′ to 3′) of the G mRNAs of RSV strains Long (subgroup A), A2 (subgroup A), and 18537 (subgroup B). The alignments were performed using the NUCALN program of Wilbur and Lipman (12). For the Long and 18537 sequences, only the nucleotides that differ from the A2 sequence are shown. Gaps introduced during alignment are represented by dashes. Gene start and end sequences (2) for the A2 strain are overlined. Initiation and termination codons are enclosed in boxes.

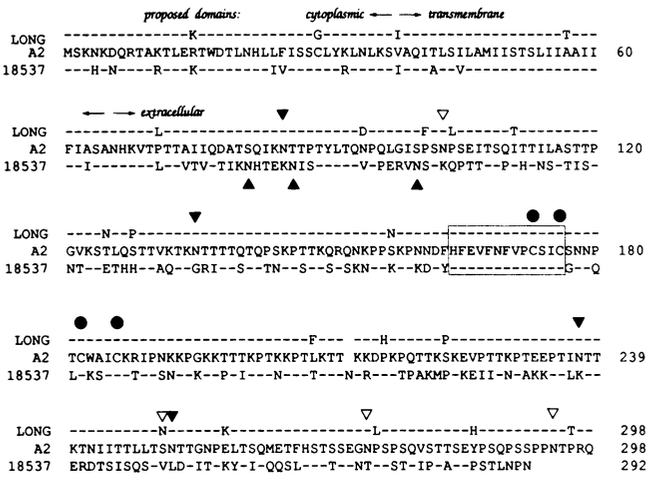

FIG. 2. Alignment of the predicted amino acid sequences (designated by the single-letter code) of the G proteins of RSV strains Long, A2, and 18537. The alignments were performed using the ALIGN program of Dayhoff (15). For the Long and 18537 sequences, only the amino acids that differ from the A2 sequence are shown; dashes above (Long) and below (18537) the A2 sequence indicate identity at these positions. A single residue gap (represented by a space) was introduced between residues 211 and 212 in the A2 and the Long sequences. Proposed cytoplasmic, transmembrane, and extracellular domains are indicated above the sequences. The positions of potential N-linked carbohydrate acceptor sites are marked by solid triangles; the four additional N-linked acceptor sites in the Long sequence that are not in the A2 sequence are marked by open triangles. The four cysteine residues in the extracellular domain are indicated by solid circles. The exactly conserved region of 13 amino acids (residues 164–176) in the extracellular domain is enclosed in a box.

been demonstrated (13, 14). Four potential acceptor sites for N-linked sugars were identified in the A2 sequence (refs. 13 and 14; Figs. 2 and 3). None of these sites was conserved in the 18537 G sequence; however, three new potential acceptor sites did occur over a 20 amino acid stretch in the NH_2-terminal portion of the proposed extracellular domain (Figs. 2 and 3). For the Long G protein, all four of the sites present in the A2 sequence were conserved. Moreover, four additional acceptor sites, all in the proposed extracellular domain (Figs. 2 and 3), were present in the Long sequence. Evidence that some of these additional sites might be used lies in

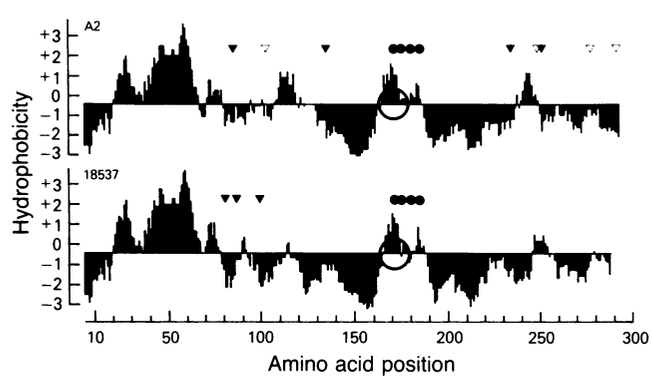

FIG. 3. Comparison of the hydropathy profiles of the G proteins of RSV strains A2 (*Upper*) and 18537 (*Lower*). The profiles were generated by using the algorithm of Kyte and Doolittle (17). Values for each position were calculated over an interval of 9 amino acids. Potential acceptor sites for N-linked carbohydrate are indicated by solid triangles; the four additional sites in the Long sequence that are not present in the A2 sequence are marked by open triangles. The positions of extracellular cysteine residues, which are exactly conserved among the three sequences, are indicated by solid circles. The exactly conserved region of 13 amino acids (positions 164–176) in the extracellular domain is enclosed in a large open circle.

previous work that demonstrated a larger effect of tunicamycin on the migration of the Long G glycoprotein than the A2 G glycoprotein (18). Thus, the locations and relative content of O-linked and N-linked carbohydrate did not appear to be conserved within or between RSV subgroups. Consistent with this, it was noted previously that the electrophoretic mobility of the G protein varied significantly among RSV strains independent of subgroup classification (19).

Another feature of the A2 G protein was a high proline content (10%), which appears to be characteristic of glycoproteins with abundant O-linked sugars (13, 14). Similarly, the 18537 G protein also had a high proline content (8.6%). Indeed, the exact position of 67% of the proline residues was conserved for the 18537 G protein, and in most instances, where prolines were not exactly conserved in the 18537 sequence, new prolines appeared nearby.

The high degree of amino acid divergence notwithstanding, the hydrophobicity plots of the unglycosylated A2 and 18537 G proteins were strikingly similar (Fig. 3). Even in regions of extreme divergence, the local hydropathic index was generally maintained. As shown in Fig. 3 and described previously (13, 14), the G protein contains a single major hydrophobic domain (residues 38–66 in both sequences) that is predicted to serve as the transmembrane anchor. The location of this hydrophobic sequence, together with the locations of potential carbohydrate acceptor sites described above, predicts that the carboxyl-terminal two-thirds of the molecule is oriented externally.

Given the overall sequence divergence between the G proteins of the A2 and 18537 strains, it was surprising that the proposed cytoplasmic and transmembrane domains exhibited extensive amino acid homology (81% and 86% identity, respectively). In sharp contrast, the sequences of the proposed extracellular domains were quite distinct (44% identity). Interestingly, however, the proposed extracellular domains of the A2, Long, and 18537 G proteins each contained a 13-residue stretch of exact amino acid identity (positions 164–176). Also, the number and positions of cysteine residues in the extracellular domains were exactly conserved; each contained four cysteine residues clustered within a single 14-amino acid region (positions 173–186) that overlapped with the exactly conserved 13-amino acid sequence. Furthermore, this highly conserved, cysteine-rich region was flanked by conserved proline residues, was relatively deficient in potential acceptor sites for O-linked carbohydrate, and was considerably removed from potential acceptor sites for N-linked carbohydrate that were clustered elsewhere in the extracellular domain.

Local secondary structure predictions by the method of Garnier *et al.* (20) showed that this highly conserved cysteine-rich region had a high probability of containing a tight turn (residues 172–181) between two β-sheets. The closely spaced cysteine residues might stabilize the predicted turn by participating in disulfide bonding. Hydropathy analysis (Fig. 3) showed that the predicted turn was moderately hydrophobic but was flanked by larger highly hydrophilic domains that might serve to orient the turn externally.

Divergence of Other 18537 mRNAs and Proteins. Complete sequences also were determined for four additional 18537 genes (the F, N, 1A, and 1C genes; unpublished data) and for several cDNAs representing polycistronic mRNAs. These additional sequences revealed the following features: (*i*) nucleotide sequences that encode protein were relatively highly conserved between the subgroups (1A, 72%; F, 82%; 1C, 83%; N, 86%); (*ii*) noncoding gene and intergenic sequences showed markedly less conservation (49% identity for 615 nucleotides); and (*iii*) amino acid sequences were relatively highly conserved (1A, 76%; 1C, 87%; F, 89%; N, 96%). Partial sequences determined for four other 18537

genes (M, P, 1B, and 22K; unpublished data) are consistent with these observations. Compared with these sequences, the nucleotide sequences encoding the G proteins (68% identity) were less well conserved. Also, the G sequences were unique among the RSV genes in that the percentage identity between the subgroups at the amino acid level (53%) was considerably less than at the nucleotide level, resulting in extensive amino acid sequence divergence in the proposed extracellular domain.

DISCUSSION

Comparison of nucleotide and predicted amino acid sequences for RSV strains representing the two major antigenic subgroups revealed extensive sequence diversity in the extracellular domain of the G glycoprotein, the presumed viral attachment protein. This sequence diversity, together with the overall sequence diversity in noncoding gene and intergenic sequences, suggest that the two RSV antigenic subgroups have evolved separately for a significant period of time.

The considerable divergence of the G mRNA and protein might have occurred as a result of selective immune pressure. In support of this possibility, (*i*) a high degree of nucleotide substitutions occurred in first or second codon positions, perhaps indicative of selective pressure for amino acid changes at these positions, and (*ii*) the vast majority of the amino acid changes occurred in the extracellular domain. The apparent lack of conservation of the positions and numbers of O-linked and N-linked carbohydrate chains in the G proteins could also contribute to antigenic differences between strains. In other viral systems, most notably influenza A virus (21), selective pressure has led to the emergence of serologically distinct strains. However, the extensive changes in the RSV attachment protein appear to have afforded only a moderate degree of antigenic heterogeneity, the significance of which is still unclear in the epidemiology of RSV-associated illness. For example, neither of the subgroups has been supplanted by the other in the 30 years since the initial isolation of RSV; in fact, both subgroups can circulate concurrently during RSV epidemics. It is possible that RSV antigenic heterogeneity might contribute to its ability to reinfect repeatedly, but this remains to be evaluated. It will be important to obtain sequences for RSV strains isolated during successive yearly epidemics and during successive infections of specific individuals to further evaluate sequence and antigenic diversity. The major neutralizing antibody response to RSV appears to be induced by the F protein (7), which is highly conserved between the RSV subgroups (89% identity; unpublished data). The antigenic differences observed between the two subgroups in *in vitro* neutralization assays is probably attributable to subgroup-specific antibody responses to the G protein; however, cross-reactive antibodies are induced by both subgroup G proteins, and, in cotton rats, significant (although incomplete) cross-subgroup protection is conferred by immunity to the subgroup A G protein (P.R.J. *et al.*, unpublished data). These cross-reactive neutralizing antibodies presumably recognize antigenic sites that are conserved between the two subgroup G proteins (see below).

Thus, selective immunologic pressure alone probably does not account for the extensive divergence of the G protein. A second, interrelated possibility is that, except for the highly conserved regions, structural or functional constraints have not greatly restricted amino acid substitutions of the RSV G protein. Consistent with this possibility, many of the amino acid substitutions were relatively nonconservative. The high mutation frequency for RNA genomes in general (22) would provide the opportunity for a high rate of nucleotide and amino acid substitution in the absence of constraints. However, some constraints on amino acid substitution must exist for the G protein, since the overall pattern of hydrophobicity and the high content of proline, serine, threonine, and carbohydrate have remained remarkably similar. With regard to other RSV proteins, stricter structural and functional constraints presumably could account for the greater conservation observed between the subgroups. For example, the substantially greater sequence conservation observed for the F protein is consistent with the view that the F proteins of paramyxoviruses have highly conserved global features of secondary structure (10). Thus, for the RSV F protein, many amino acid substitutions might not be acceptable despite selective pressure from the host immune system. In contrast, we suggest that an unusual amount of structural flexibility has permitted extensive amino acid substitution in the RSV G protein. One interesting possibility is that the extracellular domain contains a circumscribed conserved functional domain flanked by relatively divergent, nonspecific, heavily glycosylated supporting structures. An obvious candidate for a conserved functional domain, involved in receptor binding, is the conserved cysteine-containing sequence (positions 164–186) described above.

The observation that the proposed cytoplasmic and transmembrane domains of the RSV G proteins were highly conserved is in direct contrast to the divergence of these domains between serotypes of two other NH_2 terminally anchored transmembrane proteins, the paramyxovirus HN (23) and the influenza NA (24). For the HN proteins, neither domain is conserved, while for the influenza NA, the cytoplasmic tail is conserved and the transmembrane domain is not. Thus, the RSV G proteins represent a third pattern of conservation of these two regions. Intracellular processing pathways have not been well defined for proteins that contain a high content of O-linked sugars, such as the RSV G protein. The high degree of conservation of the cytoplasmic and transmembrane domains of the RSV G protein raises the possibility that these sequences contain signals that direct transport and O-linked glycosylation. Conversely, the divergence of the extracellular domains suggests an absence of specific transport or processing signals, although general features such as the high proline content might be involved. Possible roles of these domains in transport and processing could be investigated by examining chimeric proteins, expressed from chimeric cDNAs, that contain domains from the RSV G protein and from a paramyxoviral HN protein, which has a similar membrane orientation but does not contain O-linked carbohydrate.

The G proteins of the RSV subgroups A and B induce significant heterologous protection, despite extensive sequence divergence. For comparison, the relationship between antigenic cross-reactivity and sequence homology was examined for the attachment proteins of several other negative-strand viruses (Table 1). This comparison identified two general groups: group I, illustrated by the RSV G protein and the parainfluenza type 3 HN protein, was characterized by retention of antigenic cross-reactivity, despite relatively extensive sequence divergence. Group II, illustrated by the vesicular stomatitis virus G protein, the parainfluenza HN protein, and the influenza A HA was characterized by sequence divergence and a relative loss of polyclonal antibody cross-reactivity. Attachment proteins of viruses in the second group have undergone serologic divergence because of the presumed loss of important cross-reactive antigenic sites due to sequence divergence. In contrast, the RSV G proteins share antibody binding and neutralizing antigenic sites despite the overall extensive differences in amino acid homology between the subgroups.

One likely candidate for an important shared antigenic site is the previously discussed highly conserved region overlapping the four cysteine residues in the extracellular domain. Because of the suggested structural and functional impor-

Table 1. Comparison of sequence and serologic relationships of the attachment proteins of related negative strand viruses

Pairs of related negative-strand viruses	Virion attachment protein	% amino acid identity	Difference in glycoprotein-specific antibody titers,* -fold	Refs.
Group I: Sequence divergence with conservation of antigenic relatedness				
RSV				†
A2 vs. 18537	G	53	40	
A2 vs. Long		94	2	
Parainfluenza virus type 3				
Human vs. bovine	HN	75	8	25
Group II: Sequence divergence with loss of antigenic relatedness				
Parainfluenza virus				
Type 1, Sendai vs. type 3, human	HN	45	>100	26, 27
Vesicular stomatitis virus				
New Jersey vs. Indiana	G	51	>100	28, 29
Influenza virus type A‡	HA			30, 31
Subtypes				
H3 vs. H2		42	>100	
H3 vs. H1		42	>100	
H2 vs. H1		66	>100	
Drift strains				
H3/68 vs. H3/75		92	40	

*Maximum difference in antibody cross-reactivity in a glycoprotein-specific assay (see references for specific details). Assays used were as follows: RSV, ELISA; parainfluenza and influenza, hemagglutination inhibition; vesicular stomatitis virus, *in vitro* neutralization. A value of >100 indicates no detectable antibody cross-reactivity.

†Present work and P.R.J. *et al.*, unpublished data.

‡Specific subtypes were as follows: H3, A/Aichi/2/68; H2, A/JAP/305/57; H1, A/PR/8/34. Specific drift strains were as follows: H3/68, A/Hong Kong/8/68; H3/75, A/Victoria/3/75.

tance of this area, variation in the sequence (and therefore local structure) might not be tolerated. Antibodies binding within this region might neutralize infectivity by blocking virus–host cell interactions and thus may be important in cross-subgroup neutralization. The hydrophobic nature of the conserved domain might reduce its accessibility to antibodies, and sequence diversity in the surrounding hydrophilic regions could contribute to the observed incomplete cross-subgroup neutralization.

In summary, the data presented in this paper suggest a model for G protein structure and new approaches to probe the structure and function of the RSV G protein. Circumscribed sets of peptides could be used to investigate immunologic and functional properties of selected regions conserved between the two subgroups. Also, experiments involving site-specific mutagenesis could be performed to explore the structural and functional roles of precisely defined areas of the G protein. Finally, recombinant vectors expressing the subgroup B (18537) G protein could be used to complement the available data on the immunogenicity and protective efficacy of the subgroup A (A2) G protein. These data should prove valuable in the effort to develop a safe and effective RSV vaccine.

We thank Drs. Robert M. Chanock and Brian R. Murphy for support of this work and critical review of the manuscript, Ena Camargo for technical assistance, Alicia Buckler-White for synthetic oligonucleotides, and Linda Jordan for editorial assistance.

1. McIntosh, K. & Chanock, R. M. (1985) in *Virology*, ed. Fields, B. N. (Raven, New York), pp. 1285–1304.
2. Collins, P. L., Dickens, L. E., Buckler-White, A., Olmsted, R. M., Spriggs, M. K., Camargo, E. & Coelingh, K. V. W. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 4594–4598.
3. Coates, H. V., Kendrick, L. & Chanock, R. M. (1963) *Proc. Natl. Acad. Sci. USA* **112**, 958–964.
4. Anderson, L. J., Hierholzer, J. C., Tsou, C., Hendry, R. M., Fernie, B. F., Stone, Y. & McIntosh, K. (1985) *J. Infect. Dis.* **151**, 626–633.
5. Mufson, M. A., Orvell, C., Rafnar, B. & Norrby, E. (1985) *J. Gen. Virol.* **66**, 2111–2124.
6. Elango, N., Prince, G. A., Murphy, B. R., Venkatesan, S., Chanock, R. M. & Moss, B. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 1906–1910.
7. Olmsted, R. A., Elango, N., Prince, G. A., Murphy, B. R., Johnson, P. R., Moss, B., Chanock, R. M. & Collins, P. L. (1986) *Proc. Natl. Acad. Sci. USA* **83**, 7462–7466.
8. Stott, E. J., Ball, L. A., Young, K. K., Furze, J. & Wertz, G. W. (1986) *J. Virol.* **60**, 607–613.
9. Gubler, U. & Hoffman, B. J. (1983) *Gene* **25**, 263–269.
10. Spriggs, M. K., Olmsted, R. A., Venkatesan, S., Coligan, J. E. & Collins, P. L. (1986) *Virology* **152**, 241–251.
11. Collins, P. L., Huang, Y. & Wertz, G. W. (1984) *Proc. Natl. Acad. Sci. USA* **81**, 7683–7687.
12. Wilbur, W. J. & Lipman, D. G. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 726–730.
13. Wertz, G. W., Collins, P. L., Huang, Y., Gruber, L., Levine, S. & Ball, L. A. (1985) *Proc. Natl. Acad. Sci. USA* **82**, 4075–4079.
14. Satake, M., Coligan, J. E., Elango, N., Norrby, E. & Venkatesan, S. (1985) *Nucleic Acids Res.* **13**, 7795–7812.
15. Dayhoff, M. O., ed. (1978) *Atlas of Protein Sequence and Structure* (Natl. Biomed. Res. Found., Washington, DC), Vol. 5, Suppl. 3.
16. Cummings, R. D., Kornfeld, S., Schneider, W. J., Hobgood, K. K., Tolleshang, H., Brown, M. S. & Goldstein, J. L. (1983) *J. Biol. Chem.* **258**, 15261–15273.
17. Kyte, J. & Doolittle, R. F. (1982) *J. Mol. Biol.* **157**, 105–132.
18. Fernie, B. E., Dapolito, G., Cote, P. J. & Gerin, J. L. (1985) *J. Gen. Virol.* **66**, 1983–1990.
19. Norrby, E., Mufson, M. A. & Hooshmand, S. (1986) *J. Gen. Virol.* **67**, 2721–2729.
20. Garnier, J., Osguthorpe, D. J. & Robson, B. (1978) *J. Mol. Biol.* **120**, 97–120.
21. Murphy, B. R. & Webster, R. G. (1985) in *Virology*, ed. Fields, B. N. (Raven, New York), pp. 1179–1239.
22. Holland, J. J. (1984) in *Concepts in Viral Pathogenesis*, eds. Notkins, A. L. & Oldstone, M. B. A. (Springer, New York), pp. 137–143.
23. Kolakofsky, D., Curran, J., Blumberg, B., Giorgi, C., Lamb, R., Patterson, R., Hiebert, S., Venkatesan, S. & Elango, N. (1987) in *Biology of Negative Strand RNA Viruses*, eds. Mahy, B. W. J. & Kolakofsky, D. (Elsevier, Amsterdam), pp. 239–245.
24. Markoff, L. & Lai, C. (1982) *Virology* **114**, 288–297.
25. Coelingh, K. L., Winter, C. C., Murphy, B. R., Rice, J. M., Kimball, P. C., Olmsted, R. A. & Collins, P. L. (1986) *J. Virol.* **60**, 90–96.
26. Spriggs, M. K. & Collins, P. L. (1986) *J. Gen. Virol.* **67**, 2705–2719.
27. Cook, M. K. & Chanock, R. M. (1963) *Am. J. Hyg.* **77**, 150–159.
28. Gallione, C. J. & Rose, J. K. (1983) *J. Virol.* **46**, 162–169.
29. Cartwright, B. & Brown, F. (1972) *J. Gen. Virol.* **16**, 391–398.
30. Palese, P. & Young, J. F. (1982) *Science* **215**, 1468–1474.
31. Both, G. W., Sleigh, M. J., Cox, N. J. & Kendal, A. P. (1983) *J. Virol.* **48**, 52–60.